



# Improving search relevancy for oceanographic data discovery

Ed Armstrong<sup>1</sup>, Chaowei Yang<sup>2</sup>, David Moroni<sup>1</sup>, Thomas Huang<sup>1</sup>, Lewis Mcggibney<sup>1</sup>,  
Frank Greguska<sup>1</sup>, Yongyao Jiang<sup>2</sup>, Yun Li<sup>2</sup>, Christopher Finch<sup>1</sup>

<sup>1</sup>Physical Oceanographic DAAC  
NASA Jet Propulsion Laboratory, Pasadena, CA  
<sup>2</sup>George Mason University, Fairfax, VA

2018 GHRSSST19th Science Team Meeting  
Darmstadt, Germany  
7 June 2018

© 2018 All rights reserved



National Aeronautics and  
Space Administration

Jet Propulsion Laboratory  
California Institute of Technology  
Pasadena, California

# PO.DAAC Datasets – 500+ datasets



## NASA Missions & Projects

Seasat, TOPEX/Poseidon, Jason-1, NSCAT,  
SeaWinds on ADEOS-II, QuikSCAT, ISS-  
RapidSCAT, GRACE, GHRSS, SPURS,  
MEaSURES, Aquarius, CYGNSS, GRACE-FO  
(2017)



*Upcoming: COWR, AirSWOT, SWOT, GRACE-2*

## Ocean & Climate Community Driven

*Value-added datasets in support of NASA programs*

Gravity

Ocean Circulation & Currents

Ocean Surface Salinity

Ocean Surface Topography

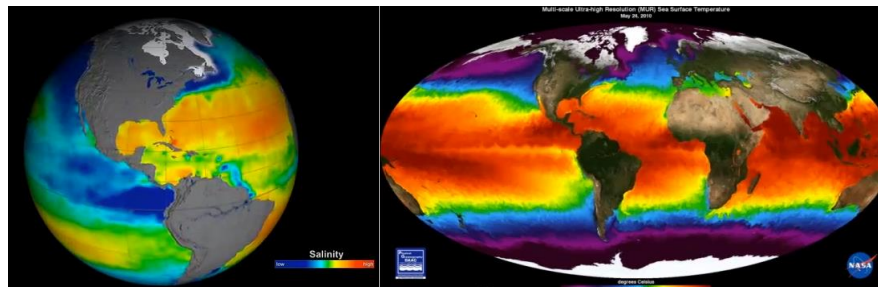
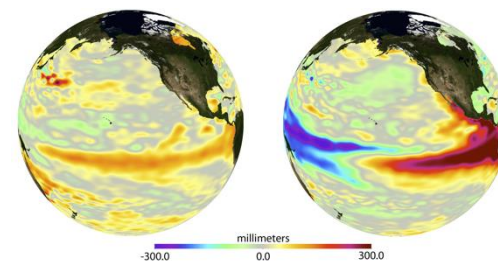
Ocean Vector Winds

Sea Surface Temperature

*Hydrology*

*Ocean Color*

*Sea Ice*





**Dataset Information Page**

- \* Information
  - \* Dataset Metadata
- \* Data Access
  - \* Direct Access
  - \* Tools and Services
  - \* Read Software
- \* Documentation
  - \* Known Issues
- \* Granule (File) Listing
- \* Citation

## Dataset Discovery

- Faceted Browsing
- Multi-level filtering
- **Keyword search**
- Dataset Information Page/DOI Landing Pages
- Granule browsing through date tree

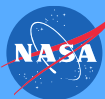
Armstrong/JPL



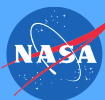
- Driven by Solr/Lucene index of PO.DAAC metadata
  - Limited search factors: term frequency (pre-defined keyword list), inverse document frequency, and dataset popularity
  - Implements a default “OR” between keywords
  - Suffers from low relevancy search precision
    - E.g., the search will often return good number of datasets (reasonable recall) but a low number of relevant datasets (precision is poor)
      - “OR” syntax often returns unrelated datasets
      - Incomplete indexing. Newer versions, release date, processing levels etc. not considered
      - User popularity (unique users) is an imperfect factor



- Ranking is a long-standing problem in geospatial data discovery...data diversity and heterogeneity, user search intent
- UWG recommendations over past several years
- *.....Improve search and discovery of PO.DAAC dataset via free text (.e.g., keyword) and facets*
- *.....Develop advanced search capabilities*
- While faceted search provides a systematic approach to group data artifacts, facets are still static and rely on manual keywords tagging.
- Search relevance requires multi-dimensional dynamic ranking of data



- **Mining and Utilizing Dataset Relevancy from Oceanographic Data (MUDROD)**
  - 2014 funded NASA Advanced Information Systems Technology (AIST) project
    - Technology Readiness Level development from an approximately Level 4 to Level 6-7
  - Specifically targeted to improve search relevance for earth science data in the PO.DAAC
    - Improving the capabilities of PO.DAAC search
  - Built on services previously implemented for the hydrology community



# Use Cases for MUDROD Development



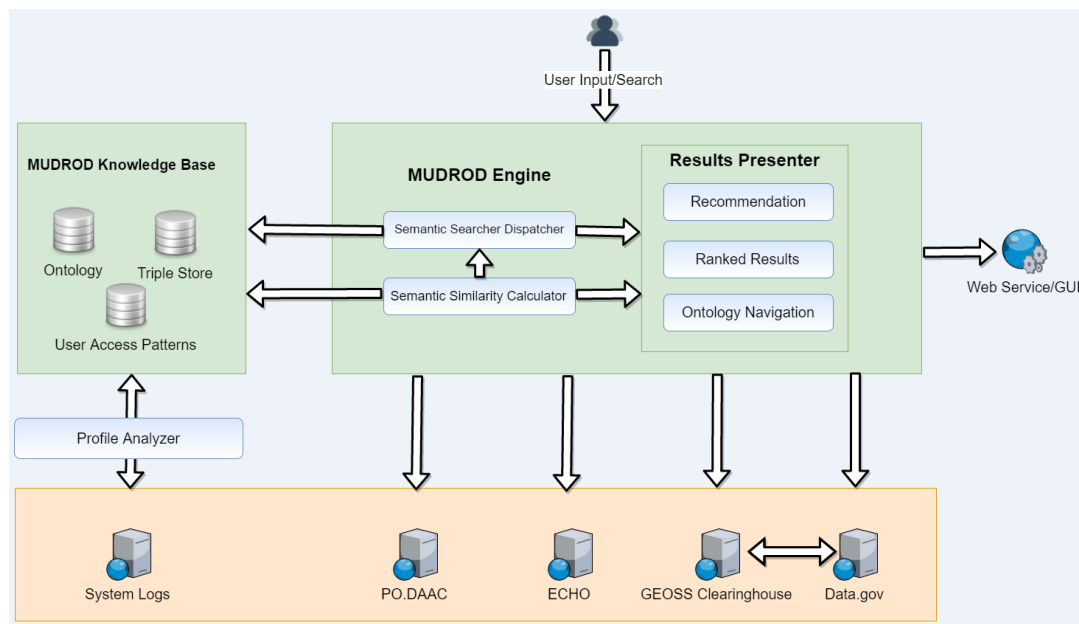
- Rank most recent versions of datasets higher
- Rank new mission dataset higher
- Allow user choice of “AND” vs “OR” or phrase keyword syntax
- Improve search across different ocean variables
- Find (and rank) related PO.DAAC datasets
- Prioritize datasets that have been vetted by “domain experts”
- Consider user search intent, e.g.
  - Climate users vs real time applications users
  - High spatial resolution vs low spatial resolution



# MUDROD search relevance methodology and technical objectives

## Objectives

- Analyze **web logs and metadata** to discover user knowledge (query and data relationships)
- Construct **knowledge base** by combining semantics and profile analyzer
- Improve data discovery by 1) better **ranking**; 2) **recommendation**; 3) **ontology navigation**





## Technology (four technological modules)

- PO.DAAC FTP and web log processing and session construction
- Semantic analysis of user queries & navigation, and metadata records
- Machine learning applied to search ranking
- Dataset recommendation engine



# Objectives and algorithm factors



- Put the most desired dataset to the top of the result list
- What **features** can represent users' search preferences for geospatial data?
- How can the ranking function reach a **balance** of all these features?
- Identified eleven features (factors) by considering user behavior, query-text match and examining common geospatial metadata attributes.
  - Geospatial metadata attributes (next slide)
  - Query – metadata content overlap (spatial similarity)
  - User behavior modeling from FTP/web logs (popularity and semantic similarity)



# Ranking features – Metadata attributes

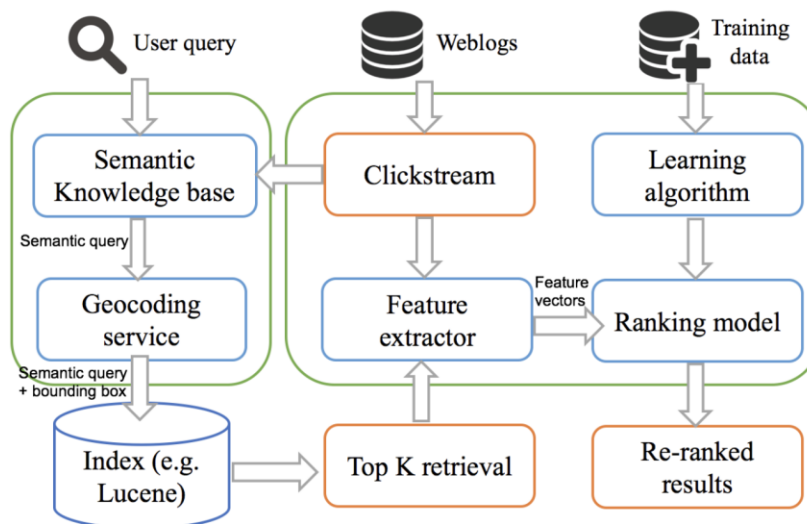
Features	Description
Release date	The date when the data were published
Processing level (PL)	The processing level of image products, ranging from level 0 to level 4.
Version number	The published version of the data
Spatial resolution	The spatial resolution of the data
Temporal resolution	The temporal resolution of the data

- Five dataset metadata features
- Verified by domains experts
- Query-independent: static, depends on the data itself, won't change with the query



# Tying it all together – Machine Learning, the Rank Support Vector Machine (RankSVM)

- One of the well-recognized Machine Learning ranking algorithms
- Convert a **ranking** problem into a **classification** problem that a regular SVM algorithm can solve
  - A classifier trained to predict the ranking order of data pairs
- A ranking problem becomes a binary classification problem, where SVM is applied to find the **optimal decision boundary**
- Has the best NCDG





# Comparison of “ocean OR wind” search results



## PO.DAAC (Solr)

### Dataset discovery

Found 382 matching dataset(s).

? Need help selecting a dataset?  
Visit the PO.DAAC Forum

Advanced search

**Free Text Search**  
Enter search text

**Temporal Search**  
Start Date  
  
Stop Date

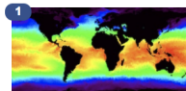
Perform Search Reset

View mode:

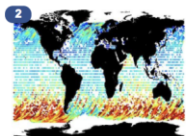


Sort By Popularity (All Time)

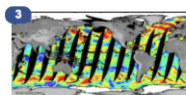
Prev 1 2 3 4 5 6 7 8 9 10 11 ... 38 39 Next



**GHRSSST Level 4 G1SST Global Foundation Sea Surface Temperature Analysis (JPL\_OUROCEAN-L4UHfnd-GLOB-G1SST)**  
**Ocean Temperature**  
Platform/Sensor: AQUA/AMSR-E , AQUA/MODIS , InSitu/InSitu ... more  
Processing Level: 4  
Longitude/Latitude Resolution: 0.01 degrees x 0.01 degrees  
Start/End Date: 2010-Jun-9 to Present  
Description: A Group for High Resolution Sea Surface Temperature (GHRSSST) Level 4 sea surface temperature analysis produced daily on an operational basis at the JPL OurOcean group using a multi-scale ... more



**TOPEX/Poseidon L2 Ocean Surface Topography Merged Geophysical Data Record Crossover ver B (TOPEX\_L2\_OST\_MGDR\_CROSSOVER)**  
**Ocean Waves, Sea Surface Topography**  
Platform/Sensor: TOPEX/POSEIDON/TOPEX ALTIMETER , TOPEX/POSEIDON/POSEIDON ALTIMETER , TOPEX/POSEIDON/TOPEX MICROWAVE RADIOMETER  
Processing Level: 2  
Along/Across Track Resolution: 11.2 km x 5.1 km  
Start/End Date: 1996-Apr-24 to 1998-Jun-26  
Description: This dataset contains the crossover points from TOPEX(ocean TOPography EXperiment)/Poseidon Merged Geophysical Data Record version B (MGDR-B). The MGDR-B combines measurements from ... more



**Cross-Calibrated Multi-Platform Ocean Surface Wind Vector L2.5 First-Look SSM/I-F14 Microwave Analyses (CCMP\_MEASURES\_ATLAS\_L3\_OW\_L2\_5\_SSMI\_F14\_WIND\_VECTORS\_FLK)**  
**Ocean Winds**  
Platform/Sensor: DMSP-F14/SSM/I  
Processing Level: 3  
Longitude/Latitude Resolution: 0.25 degrees x 0.25 degrees  
Start/End Date: 1997-May-7 to 2008-Aug-8  
Description: This dataset is derived under the Cross-Calibrated Multi-Platform (CCMP) project and contains value-added Special Sensor Microwave Imager (SSM/I) ocean surface winds from the Defense ... more



**GHRSSST Level 2P Global Skin Sea Surface Temperature from the Advanced Very High Resolution Radiometer (AVHRR) on the MetOp-A satellite produced**

## MUDROD

MUDROD Home

ocean wind

Search Operator: Or Phrase And

Showing 10 of 471 total match(es)

Default (Machine Learning Ranking)

First Previous 1 2 3 4 5 6 7 8 9 10 Next Last

**Name:** RSCAT\_LEVEL\_2B\_OWV\_CLIM\_12\_V1  
**Long Name:** RapidScat Level 2B Climate Ocean Wind Vectors in 12.5km Footprints  
**Topic:** Surface Winds  
**Platform/Sensors:** RapidScat  
**Processing Level:** 2  
**Start/End Date:** 10/03/2014 - 08/19/2016  
**Description:**  
This dataset contains the RapidScat Level 2B 12.5km Version 1.0 Climate quality ocean surface wind vectors. The Level 2B wind vectors are binned on a ... More

**Name:** ALES\_L2\_OST\_JASON2\_V1  
**Long Name:** ALES Jason-2 Coastal Altimetry Version 1  
**Topic:** Sea Surface Height, Significant Wave Height  
**Platform/Sensors:** TRSR, POSEIDON-3, AMR  
**Processing Level:** 2  
**Start/End Date:** 07/04/2008 - Present  
**Description:**  
Adaptive Leading Edge Subswathfinder (ALES) provides coastal and open ocean altimetric measurements by applying a specialized retracker to Jason-2 data. ... More

**Name:** QSCAT\_L1C\_NONSPINNING\_SIGMA0\_WINDS  
**Long Name:** QuikSCAT Level 1C Averaged Sigma-0 and Winds from Non-spinning Antenna Version 1.0  
**Topic:** Sigma Naught, Surface Winds  
**Platform/Sensors:** SEAWINDS  
**Processing Level:** 1C  
**Start/End Date:** 07/16/2010 - Present  
**Description:**  
This dataset contains geo-located and averaged Level 1B Sigma-0 measurements and wind retrievals from the SeaWinds on QuikSCAT platform, initiated in ... More

**Name:** ASCAT-B-L2-Coastal  
**Long Name:** MetOp-B ASCAT Level 2 Ocean Surface Wind Vectors Optimized for Coastal Ocean  
**Topic:** Surface Winds  
**Platform/Sensors:** ASCAT  
**Processing Level:** 2

**Related Searches**

- SURFACE WIND (1)
- WIND SPEED (0.83)
- WIND DATA (0.83)
- WIND (0.77)
- VECTOR (0.75)
- OCEAN WIND VECTOR (0.73)
- OCEAN CURRENT (0.71)
- QUIKSCAT (0.68)
- SCATTEROMETER (0.66)
- WIND VELOCITY (0.65)

Not Relevant !! (SST or SSH altimeter datasets)

- MUDROD results:
  - Recall similar
  - Precision improved !



- **Dataset heterogeneity and number, and understanding user intent still represent challenges for effective earth data search**
- MUDROD demonstrated tangible improvements in the search precision over the current default PO.DAAC Solr search result
  - Results vetted by oceanographic domain experts
- MUDROD key features:
  - Log mining to extract implicit user preferences and modeling
  - Word (metadata) similarity retrieved by data mining of user queries
  - Identification of key dataset metadata attributes as search factors
  - Implemented all 11 factors in a ML algorithm
  - A dataset recommendation algorithm implemented to improve latent data relevancy
  - The proposed architecture enables the loosely coupled software structure of a data portal and avoids the cost of replacing the existing system
    - For example the transition of technology from Elasticsearch to Solr is relatively straightforward
- Deployed at: [https://podaac.jpl.nasa.gov/podaac\\_labs](https://podaac.jpl.nasa.gov/podaac_labs) and <https://mudrod.jpl.nasa.gov>
- Publications and technical related documentation can be accessed: at <https://mudrod.github.io/>
- See Yongyao Jiang et. al, tomorrow Tuesday AM for more technical details on machine learning approach
  - IN21B-0044: *Optimizing Earth Data Search Ranking using Deep Learning and Real-time User Behavior*



# Additional potential future improvements

- Add more features (e.g., temporal similarity)
- Create training data from web logs for RankSVM
- Develop a query understanding module to better interpret user's search intent (e.g. "ocean wind level 3" -> "ocean wind" AND "level 3")
- Support near real-time data ingestion to dynamically update knowledge base
- Leverage advanced computing techniques to speed up the process



Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology